FUJITSU THE POSSIBILITIES ARE INFINITE

T2KスパコンでLinpackを測る

2008年7月25日

富士通研究所
成瀬 彰

京都大学様のT2Kスパコン

FUJITSU

- 416台の4way Quadcore Opteronサーバ
  - メモリ: 32GB (/サーバ)
  - ネットワーク: DDR-IB 4本 (/サーバ)
  - ピーク演算性能: 61.2TFLOPS
    - 416台 x 4way x 4コア x 2.3GHz x 4演算
- Linpack性能で50.5TFLOPSを記録
  - 最新のTOP500リストで34位にランクイン
    - 国内4位
  - 高い実行効率 82.5%
    - 50.5T／61.2T

Linpackとは何か

FUJITSU

- LU分解による連立1次方程式の解法
  - Ax=b　(Aは行列、xとbはベクトル)
  - Aとbがgiven、行列AをLU分解してxを求める
- TOP500のベンチマーク
  - 世界中のスパコンをLinpack性能で順位付
    - 毎年6月・11月に更新
  - 上位システムが解く問題サイズ：100～200万元
    - 演算量は100京程度
  - 高い性能を出してTOP500の上位にランクインしたい
    - サイト、ベンダ、国

TOP500のランキング

FUJITSU

- ランキングは概ねピーク演算性能で決まる
  - Linpack性能 ≒ ピーク演算性能の60～80%ぐらい
    - CPUが同じなら、台数が多い方が有利
    - 台数が同じなら、高性能なCPUが有利
  - ライト級とヘビー級を比較しても…
- 実行効率は技術次第
  - 実行効率 ＝ Linpack性能 ／ ピーク演算性能
  - 高い実行効率を出すには、HW性能を最大限に引き出す技術が必要

TOP500向けLinpack測定はレース

FUJITSU

- 一般道走行 (東京→静岡)
  - Aさん　99分
  - Bさん　100分

1分しか違わない

- カーレース (鈴鹿ラップタイム)
  - Aさん　99秒
  - Bさん　100秒

1秒も違う

- TOP500向けLinpack測定はカーレースと同じ
  - 数%の違い ＝ 大きな差

Linpack測定の動向

FUJITSU

- HPLが主流
  - High Performance Linpack Benchmark
    - A. Petitet、R. Whaleyら (テネシー大)
  - 通信ライブラリ：MPI
  - 計算ライブラリ：BLAS (or VSIPL)
    → MPIとBLASを用意すれば、実行可能
  - 多数のパラメータ (サイズ設定、アルゴリズム選択)
    → システム特性に合わせた設定が可能
- チューニングポイント
  - MPI、BLAS、HPLパラメータ

HPLの挙動 (逐次)

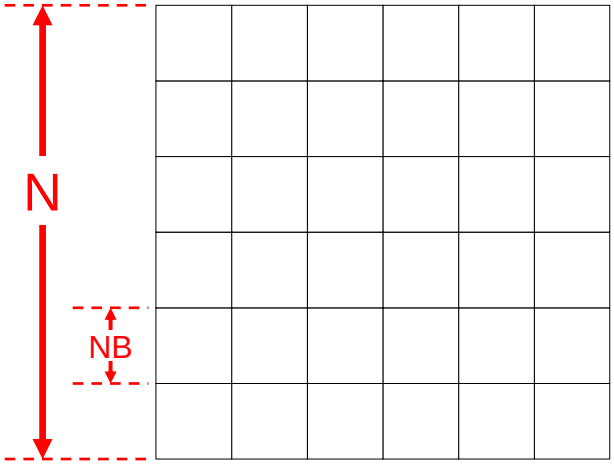
FUJITSU

- LU分解
  - 行列をブロックサイズで分割
  - 残行列がゼロになるまで、ブロックサイズ単位で下記処理を繰返す
    1. 列パネルの分解
    2. 残行列の行入替
    3. 行パネルの更新
    4. 残行列の更新
- 後退代入
  - LU分解に比べ実行時間は短い

HPLの挙動 (逐次)

FUJITSU

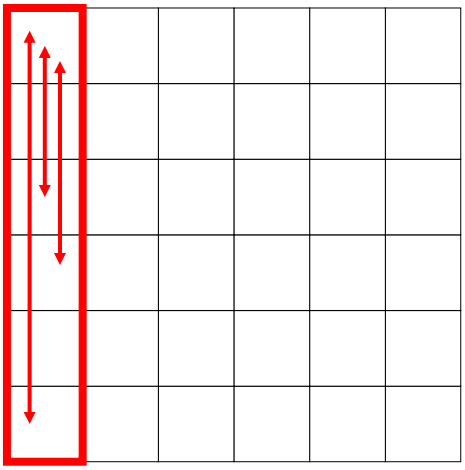
- ブロックサイズ(NB)で行列を区分



HPLの挙動 (逐次)

FUJITSU

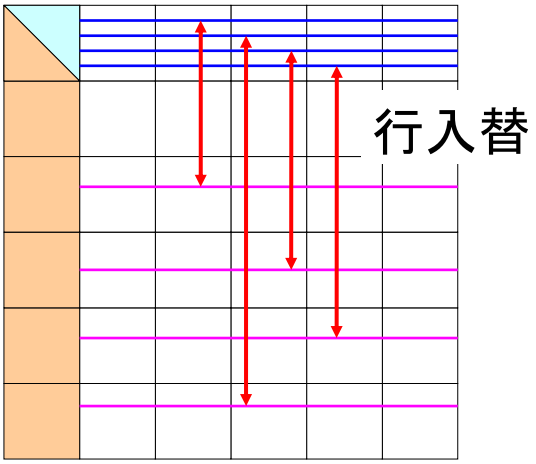
- 列パネル分解 (Panel factorization)
  - dtrsm
  - pivot



HPLの挙動 (逐次)

FUJITSU

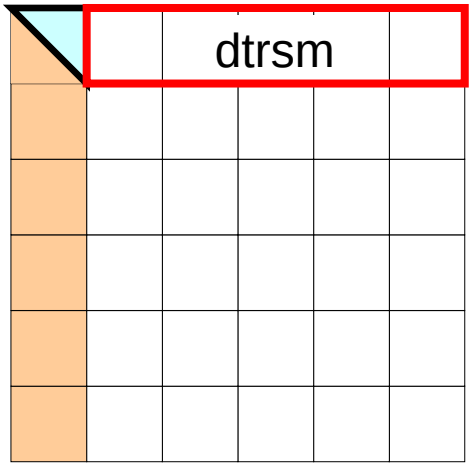
- 残行列の行入替 (Swap)
  - pivot反映



HPLの挙動 (逐次)

FUJITSU

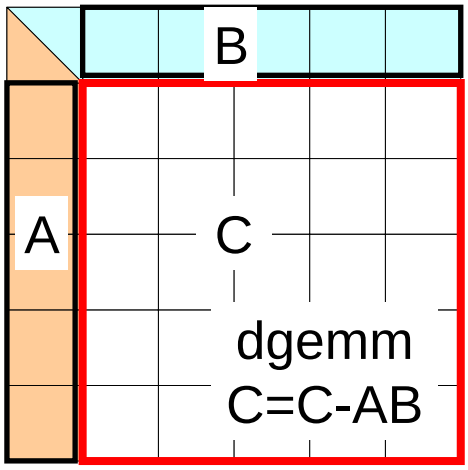
- 行パネル更新 (Update)
  - dtrsm



HPLの挙動 (逐次)

FUJITSU

- 残行列の更新 (Update)
  - dgemm
  - 計算量大



HPLで良い性能を出すポイント (逐次)

FUJITSU

- 高性能なBLAS
  - dgemm(行列積)性能はとても重要
    - dtrsm(三角行列ソルバー)性能も重要

  →行列サイズ(N)を大きくする
  Nが大きいほどdgemm性能は良くなる

  →ブロックサイズ(NB)は使用するdgemmに合わせる
  最適なNBは、使用するdgemm次第
  NBが大きいほど、高性能なdgemmを作れる

HPLの挙動 (並列)

FUJITSU

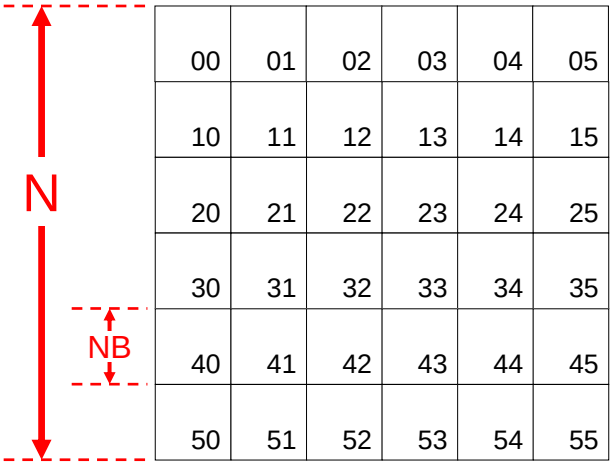
- LU分解
  - 行列をブロックサイズに分割、各プロセスに割当
  - 残行列がゼロになるまで、ブロックサイズ単位で下記処理を繰返す
    1. 列パネルの分解
    2. 列パネルのBcast　(横方向の通信)
    3. 残行列の行入替　(縦方向の通信)
    4. 行パネルの更新
    5. 行パネルのBcast　(縦方向の通信)
    6. 残行列の更新

    (*) 3.～5.、本当はまとめて処理されるが、説明簡略化のため分離

HPLの挙動 (並列)

FUJITSU

- ブロックサイズ(NB)で行列を区分

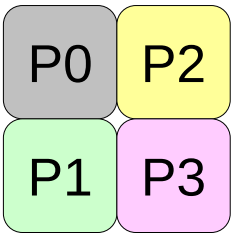


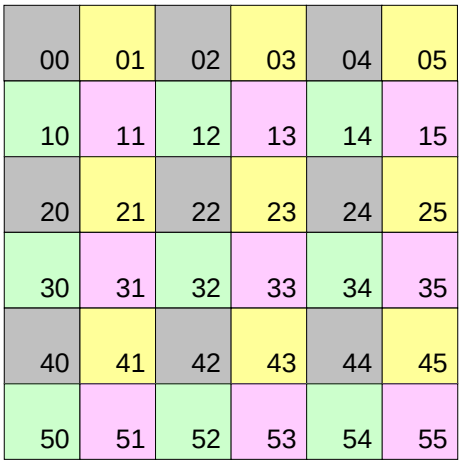
HPLの挙動 (並列)

FUJITSU

- 各ブロックをプロセスに割当

プロセス格子
(P,Q)=(2,2)

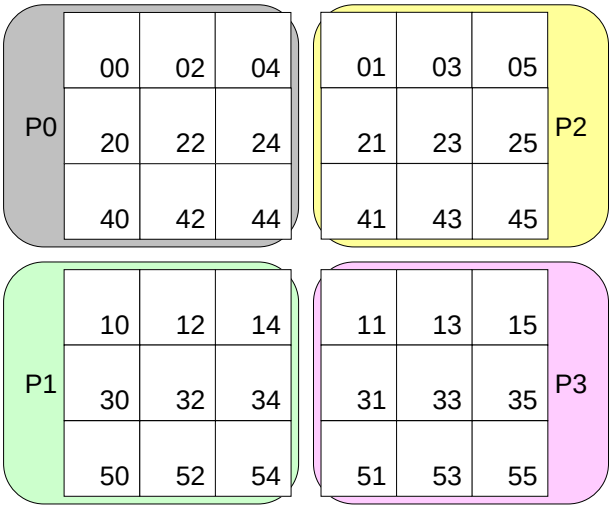




HPLの挙動 (並列)

FUJITSU

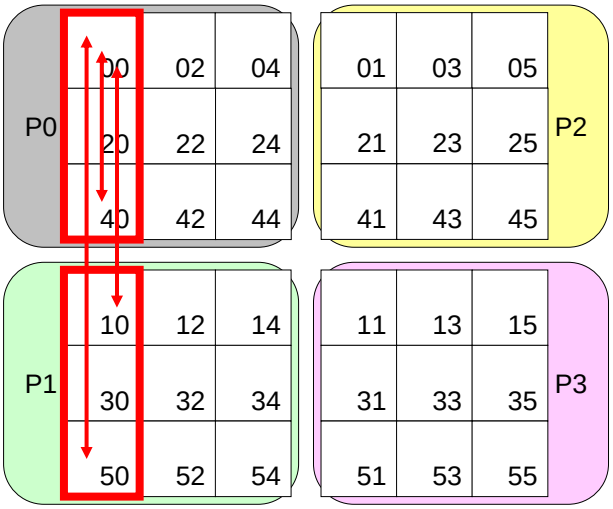
- 各ブロックをプロセスに割当



HPLの挙動 (並列)

FUJITSU

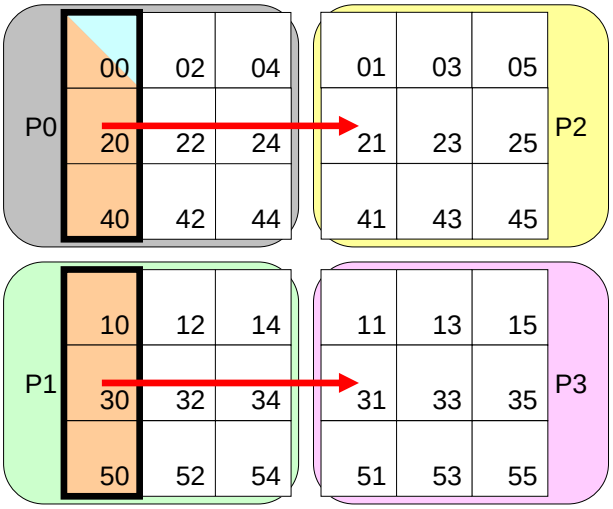
- 列パネル分解 (Panel factorization)
  - dtrsm
  - pivot
- 縦方向通信



HPLの挙動 (並列)

FUJITSU

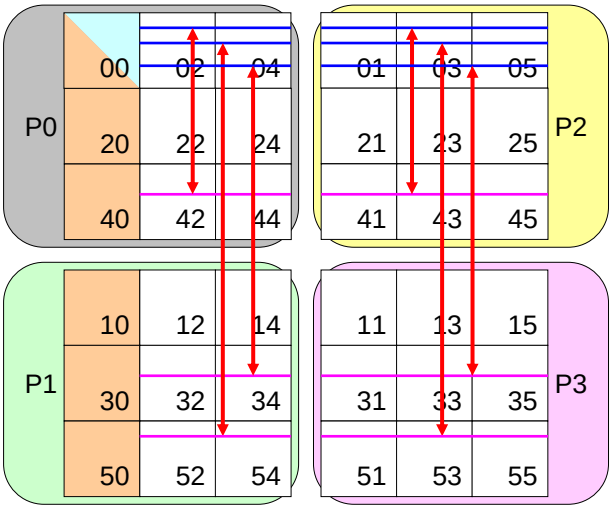
- 列パネルBcast (Panel Bcast)
- 横方向通信



HPLの挙動 (並列)

FUJITSU

- 残行列の行交換 (Swap)
  - pivot反映
- 縦方向通信



HPLの挙動 (並列)

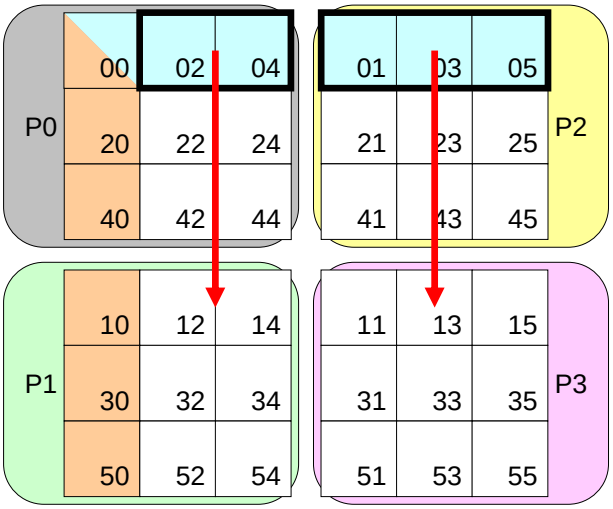
FUJITSU

- 行パネルの更新 (Update)
  - dtrsm

HPLの挙動 (並列)

FUJITSU

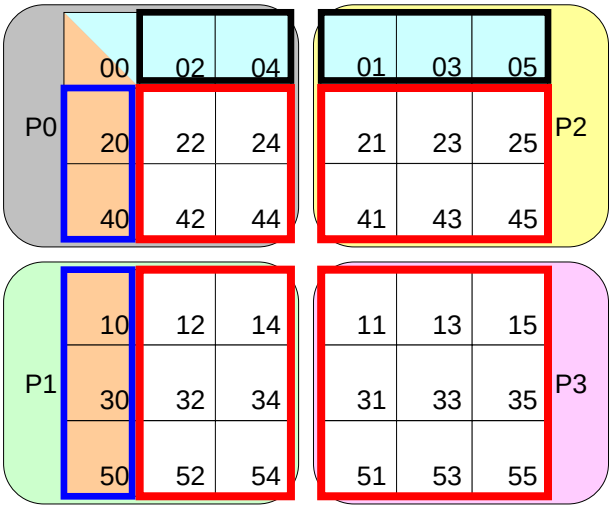
- 行パネルのBcast
- 縦方向通信



HPLの挙動 (並列)

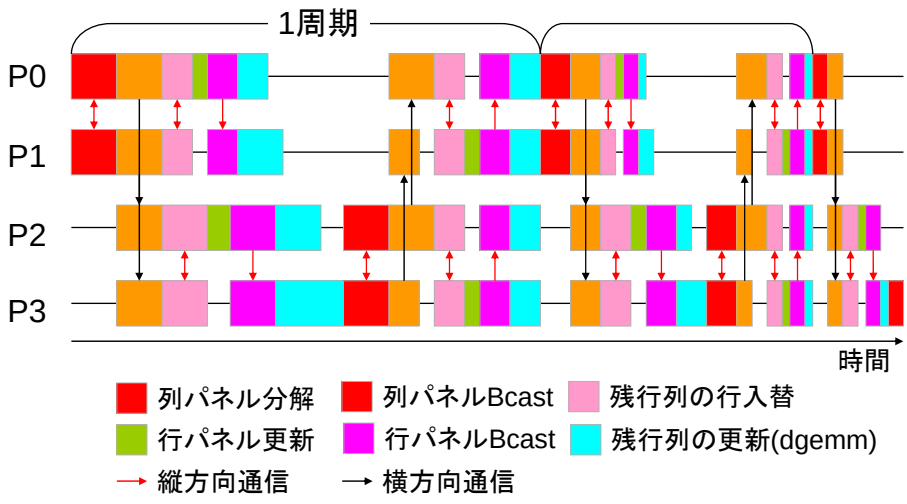
FUJITSU

- 残行列の更新
  - dgemm
- プロセス毎に計算量が違うアンバランス



HPLの挙動 (並列)

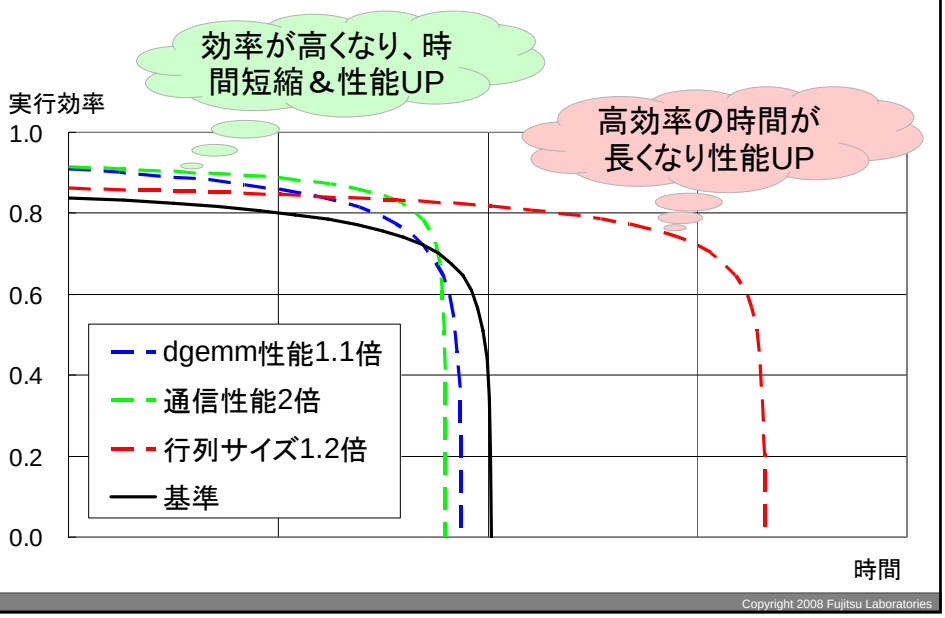
FUJITSU



- 時間が進むと、演算実行比率が低下

性能の時間変動

FUJITSU



HPLで良い性能を出すポイント (並列)

FUJITSU

- 高速な通信
  - 横方向：バンド幅が重要
  - 縦方向：バンド幅と遅延、どちらも重要
- 負荷バランス
  - プロセスの負荷(仕事量)を均等化

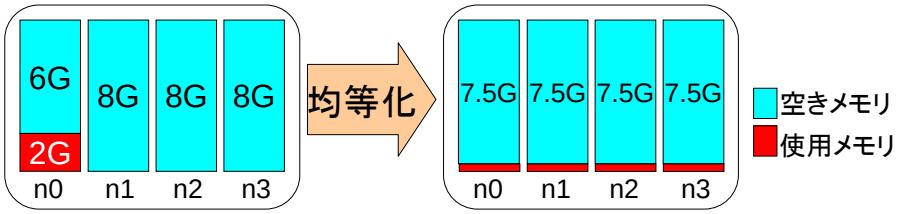
→ 行列サイズ(N)を大きくする
→ プロセス数を減らす (マルチスレッド)

[プロセス間負荷バランス]

行列サイズを大きくする

FUJITSU

- 4way QC-Opteron機は、4ノードNUMA
- 空きメモリはノード不均等
  - 使用メモリ(OS・ドライバ領域等)はノード0に偏る
  - 空きメモリ30GB、7GB使用プロセスを4つ実行(各ノード割当)
  - ノード0プロセスはリモートメモリを使用、性能低下



- 空きメモリをノード均等化
  - メモリを最大限使用してもリモートメモリアクセス発生せず
  - 性能低下の心配無く、ぎりぎりまで行列サイズを大きくできる

[プロセス間負荷バランス]

プロセス数を減らす

FUJITSU

- 京大スパコンの総コア数: 6,656コア
- Hybrid並列 (プロセス並列 x スレッド並列)
  - BLASをマルチスレッド化
  - 京大スパコンでは最大16スレッド (16コア/サーバ)

スレッド数	1	2	4	8	16
プロセス数	6,656	3,328	1,664	832	416

  - スレッド数が多くなると、性能低下
  - スレッド数が8以上になると、リモートメモリアクセスで性能低下
    - OpteronはNUMA

→ 各スレッドで性能測定、4スレッドを選択

高性能計算ライブラリと高速通信

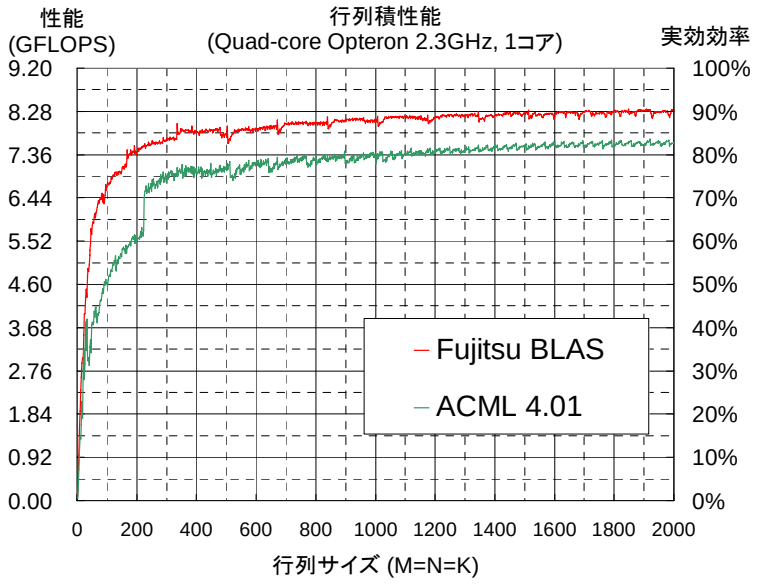
FUJITSU

- 高性能計算ライブラリ
  - Quadcore Opteron向けのBLASを開発
- 高速通信
  - InfiniBand(4系統)対応の通信ライブラリを開発

	BLAS	MPI
標準?	ACML GotoBLAS	OpenMPI MVAPICH
Fujitsu	Fujitsu BLAS	Fujitsu MPI

Fujitsu BLAS vs ACML

FUJITSU



Linpack測定

FUJITSU

- Linpack測定の割当て時間は2日
  - フル実行には7H程度必要
  - フル実行を何度も繰返す時間は無い
  - 部分実行を繰返してパラメータチューニング
- 3回のフル実行

	実行時間	TFLOPS
1回目	7.1H	48.48
2回目	6.8H	50.08
3回目	6.7H	50.51

3回目はhugetlbfs(ラージページ)を使用

試せなかったこと、失敗したこと

FUJITSU

- ラージページに最適化したdgemm
- InfiniBand(4系統)の使用方法
  - 各種トランキング数での測定
- 行列サイズ(N)の最大化
  - まだ800MB程度残っていた

まとめ

FUJITSU

- 京大T2Kスパコン、Linpack性能で50.5TFLOPS
  - 最新TOP500リストで34位 (国内4位)
  - 高い実行効率 82.5%
- あと2～3日欲しかった..

FUJITSU

THE POSSIBILITIES ARE INFINITE